ENGN 2520 / CSCI 1950-F Homework 7
Due Monday April 29 by 4pm

In this homework you will implement the EM algorithm for fitting mixtures of Gaussians.

As described in class EM is an iterative method with two steps. We start with some initial parameters and repeatedly update them with two steps:

1. Given the current parameters, compute the probabilities that each data point comes from each mixture component.

2. Compute new parameters using the membership probabilities from step 1.

For debugging purposes you should check that each iteration of EM is increasing the log-likelihood by looking at a plot of the log-likelihood over time.

In practice you should run this whole process multiple times (say 10) from several different initial parameters, and select the final parameters that lead to the highest log-likelihood. This is important because EM will find a local maxima of the likelihood function.

One common approach is to select initial parameters for a mixture of K gaussians by randomly selecting K data points to define the initial means, and setting the initial covariances to a multiple of the overall data covariance. Another common approach is to randomly assign initial cluster memberships to the data points and calculate initial parameters based on this initial assignment.

You will have to decide when to stop the algorithm. For example you might decide to stop if the increase in log-likelihood between two steps is sufficiently small.

Recall that the likelihood of a mixture of Gaussians can go to infinity if the entries in the diagonal of a covariance matrix goes to zero. You need to prevent this from happening! One simple approach is to always check if a diagonal entry in an estimated covariance is below a small threshold, and if so replace it with the treshold. A more principled solution is to put a prior on the model parameters.

The class website has two 2-dimensional datasets. One is from a mixture of 2 Gaussians and another from a mixture of 3 Gaussians. You should run your algorithm on these datasets and turn in:

1. A plot of the log-likelihood over time for the best choice of initial parameters (leading to the best final log-likelihood).

2. The resulting parameters of the mixture models.

3. A visualization of the estimated means and covariances over the datasets. In each case you should make a plot showing the data points and the estimated Gaussians by plotting their means and ellipsoids that show the estimated covariances.

4. Your code.